UNIVERSITI PUTRA MALAYSIA

HIERARCHICAL BAYESIAN SPATIAL MODELS FOR DISEASE MORTALITY RATES

RAFIDA MOHAMED ELOBAID

HIERARCHICAL BAYESIAN SPATIAL MODELS FOR DISEASE MORTALITY RATES

By

RAFIDA MOHAMED ELOBAID

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

February 2009

To
My parents,
Jailani, Lina, Mohamed and Munzir

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

HIERARCHICAL BAYESIAN SPATIAL MODELS FOR DISEASE MORTALITY RATES

By

RAFIDA MOHAMED ELOBAID
February 2009

Chairman : Associate Professor Noor Akma Ibrahim, Ph.D

Institute : Institute for Mathematical Research

The spatial epidemiology is the study of the occurrences of a disease in spatial locations. In spatial epidemiology, the disease to be examined usually occurs within a map that needs spatial statistical methods to model the observed data. The methods used should be appropriate and catered for the variation of the disease. The classical approach, which used to estimate the risk associated with the spread of the disease, did not seem to give a good estimation when there were different factors expected to influence the spread of the disease.

In this research, the relative risk heterogeneity was investigated, while the hierarchical Bayesian models with different sources of heterogeneity were proposed using the Bayesian approach within the Markov Chain Monte Carlo

(MCMC) method. The Bayesian models were developed in such a way that they allowed several factors, classified as fixed and random effects, to be included in the models. The effects were the covariate effects, interregional variability and the spatial variability, which were all investigated in three different hierarchical Bayesian models. These factors showed substantial effects in the relative risk estimation.

The Bayesian approach, within the MCMC method, produced stable estimates for each individual (e.g. county) in the spatially arranged regions. It also allowed for unexplained heterogeneity to be investigated in the disease maps. The disease maps were employed to exploratory investigate the spread of the disease and to clean the maps off the extra noise via the Bayesian approach to expose the underlying structure.

Using the MCMC method, particular sets of prior densities over the space of possible relative risks parameters and hyper-parameters were adopted for each model. The products of the likelihood and the prior densities produced the joint and conditional posterior densities of the parameters, from which all statistical inferences can be made for each model. Convergence of the MCMC simulation to the stationary posterior distributions was assessed. This was achieved by monitoring the samples of the history graphs for posterior means of the

parameters, applying statistical diagnostic test and conducting sensitivity analysis for several trials of different choices of priors.

The hierarchical models and the classical approach were applied on a spatial set of lip cancer data. The spatial correlation among the counties was examined and found to be spatially correlated. The results of the estimated relative risk for each county were compared with the result of the maximum likelihood estimation using the disease maps.

The final model selection was accomplished by applying the deviance information criterion. The performance of each model was investigated using the posterior predictive simulations. The predictive simulation for each model was carried out using the Bayesian analysis results of the real data. The graphical and numerical posterior predictive checks were used as the assessment tests for each model. The numerical results showed a good agreement with the graphical results, in which the full model with both fixed and random effects was appropriate since it was found to be capable of providing the most similar values of the original and predicted samples compared to the other models. This model was also found to be flexible since it can be reduced or extended according to the nature of the data. Nevertheless, great care must be considered in the choice of prior densities.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk Ijazah Doktor Falsafah

PEMODELAN BAYESAN BERHIERARKI RERUANG BAGI KADAR KEMATIAN PENYAKIT

Oleh

RAFIDA MOHAMED ELOBAID

Februari 2009

Pengerusi : Profesor Madya Noor Akma Ibrahim, Ph.D

Institut : Institut Penyelidikan Matematik

Epidemiologi reruang adalah kajian terhadap berlakunya sesuatu penyakit dalam lokasi reruang. Dalam epidemiologi reruang, penyakit yang perlu dikaji biasanya berlaku di persekitaran sesuatu peta yang memerlukan kaedah berstatistik untuk memodelkan data yang tercerap. Kaedah yang digunakan perlulah sesuai dan boleh menampung kepelbagaian penyakit. Pendekatan klasik, yang digunakan untuk menganggar risiko terhadap penularan penyakit seolah-olah tidak memberi suatu anggaran yang baik apabila terdapat pelbagai faktor yang dijangka akan mempengaruhi merebaknya penyakit.

Dalam kajian ini, risiko keheterogenan relatif diselidiki dan model hierarki Bayesan dengan punca keheterogenan berbeza dicadangkan menggunakan pendekatan Bayesan di dalam kaedah Rangkaian Markov Monte Carlo

(MCMC). Model Bayesan dibangunkan sedemikian rupa supaya ianya dapat mengambil kira beberapa faktor, diklasifikasi sebagai kesan tetap dan rawak ke dalam model. Kesan yang diambil kira adalah kesan kovariat, kesan ubahan di antara kawasan dan kesan ubahan reruang. Kesemuanya dikaji dalam tiga model Bayesan Hierarki yang berbeza. Faktor ini memperlihatkan terdapat kesan yang besar dalam anggaran risiko relatif.

Pendekatan Bayesan, di dalam kaedah MCMC menghasilkan anggaran yang stabil bagi setiap individu (cth. daerah) dalam kawasan reruang teratur. Ia juga mampu untuk mengkaji selanjutnya keheterogenan yang tidak boleh diterangkan dalam peta penyakit. Peta penyakit digunakan untuk menjalankan kajian jelajahan terhadap sebaran penyakit dan membersihkan peta daripada ganguan berlebihan melalui pendekatan Bayesan untuk memperlihatkan struktur sebenar data.

Menggunakan kaedah MCMC, set ketumpatan prior tertentu atas ruang parameter risiko relatif dan parameter-hyper yang mungkin telah digunapakai bagi setiap model. Hasil darab ketumpatan kebolehjadian dengan prior menghasilkan ketumpatan posterior tercantum, bersyarat dan marginal bagi parameter yang membolehkan pentakbiran statistik dilaksanakan bagi setiap model. Titik penumpuan simulasi MCMC kepada taburan posterior pegun

dinilai dengan memantau sampel dari peta sejarah bagi parameter min posterior.

Model berhierarki dan pendekatan klasik diterapkan dengan menggunakan set data reruang kanser bibir. Korelasi reruang di antara daerah dikaji dan didapati ianya berkorelasi. Keputusan daripada risiko relatif yang dianggarkan bagi setiap daerah dibandingkan dengan keputusan anggaran kebolehjadian maksimum melalui peta penyakit.

Pemilihan model yang muktamad dilaksanakan menggunakan kriteria informasi devian. Prestasi setiap model diselidiki menggunakan simulasi posterior ramalan. Simulasi ramalan bagi setiap model dijalankan menggunakan keputusan analisis Bayesan dari data sebenar. Pemeriksaan posterior ramalan secara bergraf dan berangka digunakan untuk menilai setiap model. Keputusan berangka menunjukkan kesamaan dengan keputusan bergraf yang mana model penuh dengan kedua-dua kesan tetap dan rawak adalah sesuai disebabkan ia mampu memberikan keputusan yang nilainya hampir sama dengan model sebenar dan ramalan berbanding dengan model yang lain. Model ini juga didapati fleksibel kerana ia boleh dikecilkan atau dikembangkan mengikut keadaan data. Walau bagaimanapun pemilihan yang teliti adalah perlu apabila menentukan ketumpatan prior.

ACKNOWLEDGEMENTS

First and foremost, all praised to the Almighty Allah (S.W.A) who has been very kind for giving me the opportunity to pursue the PhD programme.

I wish to express my sincere appreciation and heartiest gratitude to my supervisor, Associate Prof. Dr. Noor Akma Ibrahim, for her thoughtful comments, valuable guidance and supervision in preparing this thesis. My thanks also goes to my committee member, Associate Prof. Dr. Isa Bin Daud, for his valuable comments, suggestions and critical review in the course of writing my thesis. I express my sincere gratefulness and heartiest appreciation to my committee member, Dr. Mahendran Shitan, for his comments on the analysis conducted throughout the preparation of this thesis.

I wish to thank my course lecturers at the Institute for Mathematical Research (INSPEM) and the Dept. of Mathematics, Faculty of Science, Universiti Putra Malaysia, for their valuable assistance during the study, without which my study might not have been possible.

I want to express my deepest gratitude to my beloved husband, Jailani, who has always urged me to undertake higher education. I am extremely grateful to him

for his patience, sacrifices and support, throughout the period of my studies, and without his support, my studies might not have been possible. I also wish to express my deepest gratitude to my wonderful parents who have encouraged me to undertake higher education and always supported me. I would like to convey my deepest sense of appreciation to my sisters, particularly Manal, and to my brother for their prayers and encouragement; not to forget my sister-in-law, Makarim, for extending her support during my studies. My special thanks also go to my daughter, Lina, and my sons, Mohamed and Munzir, who have sacrificed a lot in the course of my studies which can never be repaid.

I certify that an Examination Committee has met on ---------- to conduct the final examination of Rafida Mohamed Elobaid on her Doctor of Philosophy thesis entitled, "Hierarchical Bayesian Spatial Models for Disease Mortality Rates**",** in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations, 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows:

Malik bin Hj. Abu Hassan, PhD
Professor
Institute for Mathematical Research
Universiti Putra Malaysia
(Chairman)

Kassim Haron, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Mohd. Rizam Abu Bakar, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Abdul Aziz Jemain, PhD
Professor
Faculty of Science and Technology
Universiti Kebangsaan Malaysia
Malaysia
(External Examiner)

BUJANG KIM HUAT, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 28 April 2009

This thesis was submitted to the Senate of Universiti Putra Malaysia, and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Noor Akma Ibrahim, Ph.D
Associate Professor
Institute for Mathematical Research
Universiti Putra Malaysia
(Chairman)

Isa Bin Daud, Ph.D
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

Mahendran Shitan, Ph.D
Faculty of Science
Universiti Putra Malaysia
(Member)

HASANAH MOHD. GHAZALI, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: **14 May 2009**

DECLARATION

I declare that the thesis is my original work, except for the quotations and citations which have been duly acknowledged. I also declare that it has not been previously and is not concurrently submitted for any other degree at Universiti Putra Malaysia or any other institutions.

_____

Rafida Mohamed Elobaid

Date:……..

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
AREB	Absolute Relative Estimated Bias
BF	Bayes Factor
BIC	Bayesian Information Criterion
CAR	Conditional Autoregressive Model
CI	Confidence Interval / Credible Interval
CVM	Correlated Variability Model
DIC	Deviance Information Criterion
EB	Estimated Bias
EM	Expectation Maximization
ERMSE	Estimated Root Mean Square Errors
ESE	Estimated Standard Error
GIS	Geographical Information System
GSM	Global Spatial Model
ICAR	Intrinsic Conditional Autoregressive
IVM	Interregional Variability Model
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimation
MV	Moving Average Model

NIC	Network Information Criterion
pdf	Probability density function
RR	Relative Risk
SAR	Simultaneous Autoregressive model
sd	Standard error
SMR	Standardized Mortality/Morbidity Rate
TIC	Takeuchi Information Criterion

CHAPTER 1

OVERVIEW OF THE STUDY

1.1 Introduction

Statistical studies play an important tool in scientific discovery, policy formulation and business decisions. Applications of statistics are ubiquitous that include clinical decision making, conducting an environmental risk assessment, setting insurance rate, etc.

Statistics defined as the discipline which concerns with the treatment of numerical data derived from groups of individuals. These individuals often include people, like those suffering from a certain disease, or those living in a certain area. They may also be animals or other organisms.

Statistical analysis of epidemiology has become a topic of considerable interest to statisticians and researchers in areas such as medical, biological and ecological sciences, public health, as well as environmental and geographical studies. They are usually concerned about drawing conclusions from numerical data, and about quantities which are not observed. These statistical conclusions are usually called statistical inferences.

Epidemiology is the study of how often diseases occur in different groups of people and why. It can also be defined as the study of the occurrence of diseases in relation to their explanatory factors. A key feature of epidemiology is the measurement of disease outcomes in relation to a population at risk. The population at risk is the group of people, healthy or sick, who are counted as cases if they have the disease being studied. Epidemiological information is used to plan and evaluate strategies to prevent an illness, and it also serves as a guide to the management of patients, in whom this particular disease has already developed.

Spatial epidemiology is the study of the occurrences of disease in spatial locations along with the disease explanatory factors. In spatial epidemiology, the disease to be examined usually occurs within a particular map, and the data are expressed as a point location (case event). The data can also be aggregated as a count of the disease within a sub-region of the map. Both data types need spatial statistical methods to model the observed data. The methods used should be appropriate and catered for the variation of the disease (i.e. which is generated from the population at risk of the disease).

Advances in statistical methodology, geographic information systems, and the availability of geographically referenced health and environmental data, have created new opportunities to investigate the variation of diseases. However,